ZAMBIA INFORMATION COMMUNICATION TECHNOLOGY (ICT) JOURNAL

Volume 1 (Issue 1) (2017) Pages 41-46

USE OF DATA MINING TECHNIQUES IN HUMAN RESOURCE MANAGEMENT

Okaile R. Marumo
University of Botswana
Computer Science Department
Gaborone, Botswana
rodneymarumo@gmail.com

Tumisang Angela Mmopelwa
University of Botswana
Computer Science Department
Gaborone, Botswana
ammopelwa@gmail.com

Abstract— *In the past few years, Analytics has rapidly risen in among organizations within the field of human resource management. To the present date, however, Human Resource Analytics has not been subject to a lot of scrutiny from educational researchers. The aim of this paper is so to look at Different Mining Techniques could be implemented in the HR Department to enhance or support their decision making process. This will improve existing practices of HR analytics and will deliver transformational change indeed.*

Keywords—Big Data, Data Science, HR Analytics , predictive analytics, high-potential identification

I. Introduction

In current highly competitive environment, talented people are the most valuable assets. During last years, large investments were put into tools and information systems to manage performance, hiring, compliance and employees' development to enhance its capabilities and increase effectivity.

Using data produced by these tools and systems typically implemented into enterprise HR departments, most companies can provide reports at least at some basic level. Organizations that already launched digital transformation processes do take things one step further by accompanying their reporting with basic analysis of HR metrics.

They are usually able to go through data from several previous periods to assess positive or negative trends, or to create benchmarks comparing their performance against their competitors across time and regions. However, to bring real value and help driving the business competitiveness, HR analytics utilization needs to go far beyond.

The biggest struggles in achieving better utilization of data resources and information systems are inefficient use of the data, asking wrong questions and lack of analytical ability in HR environment in general. HR departments are in need for analytically capable people enabled to provide right insights combining reporting skills and domain knowledge. J. E. Newman [9] outlined this combination of right analytical approach and experience is the crucial premise for successful HR IS and data utilization. O this paper will demonstrate how Data Mining and Big Data techniques can be applied in Prediction and understanding the attrition of employees the case study of Botswana. To avoid a huge loss and wide influence, a company has no alternative but to decrease or slow its employee's turnover rate and to find out the true causes for employees' turnover. As a result, to assist companies in building an early warning system of predicting their employees' leaving, this paper will investigate the causes through a hybrid feature selection model

A. *Turnover Intention*

To understand why industries staff, leave their jobs, Firstly, the definition of 'turnover' ought to be clearly outlined, which we simply describe as job leaving. It includes that which transfer their job from one location to another (special transfer) highlighted A. C. Bluedron [7], or from one job to another (industrial transfer), or from one industry to another (industrial transfer).

It could also happen from one organization to another (e.g., industry and organization) as described by J. E. Newman [9], input and output of employee. Narrative definition of employee turnover means that people quit their current job. Price [3] defined 'employee turnover' as that personal move over the boundary of the other organization, and this could mean entering to or leaving an organization. Ferratt and Short [8,9] outlined it as break off relationship between staff and, despite who caused this it will also be called employee turnover.

Traditionally, there are two types of turnover Carlo Dell [3] explains: (1) Voluntary turnover: Most of job movement is addressed by employees, a personal choice of employee turnover,(2) Involuntary turnover the employee turnover is not controllable, such as retire, dismiss or death. These circumstances, although labor addresses it, should not be viewed as employee turnover, because the employees have no choice.

Turnover intention (TOI) is the best factor for predicting turnover [10]. Turnover intention means the strength of intention an individual must leave his present job and look for another job opportunity [11]. Many studies show that employee turnover intention has strong relation to organizations [13]. Employees have an intention or plan to leave their jobs because of the work they are doing and the organization they are in.

B. *Feature Engineering*

Before feature selection, a set of training set is given, where each set is represented by n number of features and an output label. Many pattern classifications have been investigated to classify new, unseen instances based on extracting useful knowledge from the training set. In Theory, all features of each instance will be considered for classification. But in practice, classification tasks usually contains insignificant or redundant features, which may degrade the classification accuracy. Consequently, many feature subset selection approaches [3,4,7] have been developed to reduce the dimensionality in pattern classification. In other words, irrelevant or redundant features will be removed from the original feature set

Feature subset selection is a process that selects important or relevant features from the original feature set. It is also a search problem [6], where each search state in the search space identifies a possible feature subset. Feature subset selection offers many advantages for pattern classification. Firstly, the cost of gathering training or unseen instances can be reduced. Secondly, pattern classification models can be constructed faster. The accuracy and the ability of the learning models can be improved. If each set contains n features, the search space will contain 2n candidate feature subsets. On the other hand Exhaustive search through the entire search space has a very costly when it comes to computing and thus is usually unfeasible in practice, even for medium-sized n [4]. It is also difficult to select a best feature subset for pattern classification from the whole search space feature selection is a process of choosing feature subset from feature sets [7, 12]. To make sure the feature Subset is optimal, a specific subset evaluation is necessary.

II. Evaluation of Feature Selection Method

In this paper, Random forest was used to perform feature selection because. Random forests are among the most popular machine learning methods thanks to their relatively good accuracy, robustness and ease of use. They also provide two straightforward methods for feature selection: mean decrease impurity and mean decrease accuracy. Random forest consists of many decision trees. Every node in the decision trees is a condition on a single feature, designed to split the dataset into two so that similar response values end up in the same set. The measure based on which the (locally) optimal condition is chosen is called impurity. For classification, it is typically impurity or information gain/entropy and for regression trees it is variance [8]. Thus, when training a tree, it can be computed how much each feature decreases the weighted impurity in a tree. For a forest, the impurity decreases from each

Feature can be averaged and the features are ranked according to this measure.

A. *Selecting Principal Components*

To decide which eigenvector(s) can dropped without losing too much information for the construction of lower-dimensional subspace, we need to inspect the corresponding eigenvalues: The eigenvectors with the lowest eigenvalues bear the least information about the distribution of the data; those are the ones can be dropped

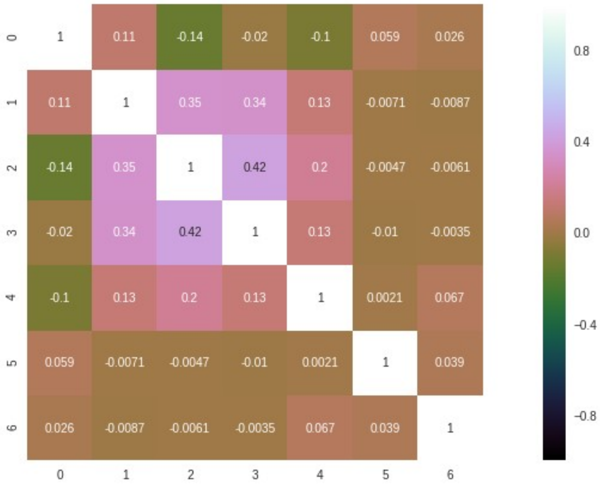


Figure 1: Correlation between features

Key:

0 – Satisfaction Level　　**1**–Last Evaluation

2 – Number of Projects　　**3**–Average Monthly Hours

4–Time Spent in Company　　**5**– Work Accidents

6 – Promotion last 5 years

III. Comparison of Classication Techniques

Three classifiers were evaluated in this paper Logistic Regression, Random Forest and Logistic Regression, sometimes referred to as the sigmoid function was developed by means of statisticians to describe residences of populace increase in ecology, rising fast and maxing out at the carrying capability of the surroundings. It’s an S-Shaped curve that may take any actual-valued range and map it right into a value between 0 and 1, however in no way exactly at these limits.

$$\frac{1}{1 + e^{(-z)}}$$

Equation 1 : Sigmoid Function

A random forest classifier is a Meta estimator that suits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to enhance the predictive accuracy and minimize over-fitting. The sub-pattern size is continually the same as the unique input pattern length however the samples are drawn with alternative if bootstrap=actual.

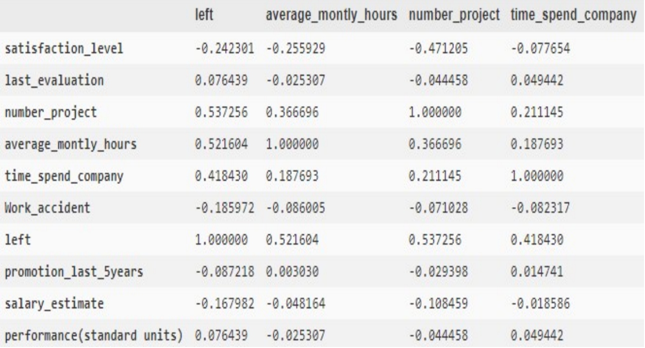
IV. DATASET

The Dataset used in this experiment to predict which valuable employees will leave next was obtained from the Human Resource Development Counsel of Botswana from a survey conducted from 1st of November 2016 to 31st of June 2017 of Botswana, the dataset has a total of 7851 employees from different sectors i.e. IT, Management, Support, Marketing, Sales etc. . Fields in the dataset include:

Satisfaction Level
Last evaluation
Number of projects
Average monthly hours
Time spent at the company
Whether they have had a work accident
Whether they have had a promotion in the last 5 years

V. RESULTS ANALYSIS

To examine the accuracy of the two methods, the investigator had a test on the prediction of turnover classification by using collected data of 7581 employees various fields. There were 44 features selected. These features were considered as feature variables and analyzed based on Random Forest feature selection method

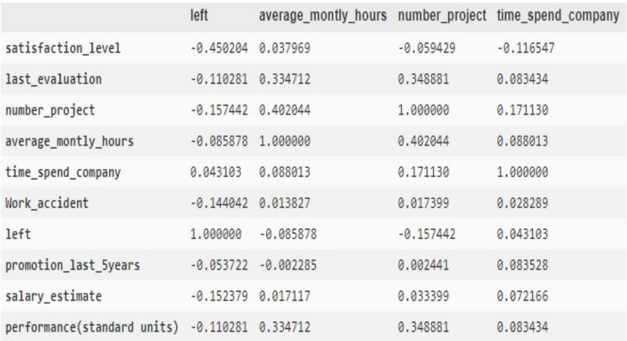


	left	average_montly_hours	number_project	time_spend_company
satisfaction_level	-0.242301	-0.255929	-0.471205	-0.077654
last_evaluation	0.076439	-0.025307	-0.044458	0.049442
number_project	0.537256	0.366696	1.000000	0.211145
average_montly_hours	0.521604	1.000000	0.366696	0.187693
time_spend_company	0.418430	0.187693	0.211145	1.000000
Work_accident	-0.185972	-0.086005	-0.071028	-0.082317
left	1.000000	0.521604	0.537256	0.418430
promotion_last_5years	-0.087218	0.003030	-0.029398	0.014741
salary_estimate	-0.167982	-0.048164	-0.108459	-0.018586
performance(standard units)	0.076439	-0.025307	-0.044458	0.049442

Table 1: Correlations for Above Average Performers

For Above Average Performers: we see leaving the company is highly correlated with working many hours, taking on many projects and time spent with the company, Monthly hours and number of projects is highly correlated with a high level of dissatisfaction whilst Time spent with the company also seemed to correlate with number of projects taken on At the same time, there was no correlation between number of projects and promotions or higher performance evaluations

Table 2: Correlations for Below Average Performers



	left	average_montly_hours	number_project	time_spend_company
satisfaction_level	-0.450204	0.037969	-0.059429	-0.116547
last_evaluation	-0.110281	0.334712	0.348881	0.083434
number_project	-0.157442	0.402044	1.000000	0.171130
average_montly_hours	-0.085878	1.000000	0.402044	0.088013
time_spend_company	0.043103	0.088013	0.171130	1.000000
Work_accident	-0.144042	0.013827	0.017399	0.028289
left	1.000000	-0.085878	-0.157442	0.043103
promotion_last_5years	-0.053722	-0.002285	0.002441	0.083528
salary_estimate	-0.152379	0.017117	0.033399	0.072166
performance(standard units)	-0.110281	0.334712	0.348881	0.083434

For Average/Below Average Performers: it indicates that leaving the company is highly correlated satisfaction level. No other metric is as correlated for this group, as opposed to the Above Average group, Monthly hours and number of projects is highly correlated with higher performance reviews.

To further explore why good people leave their post we took went deep to the departments view and explore below we the findings per department per view

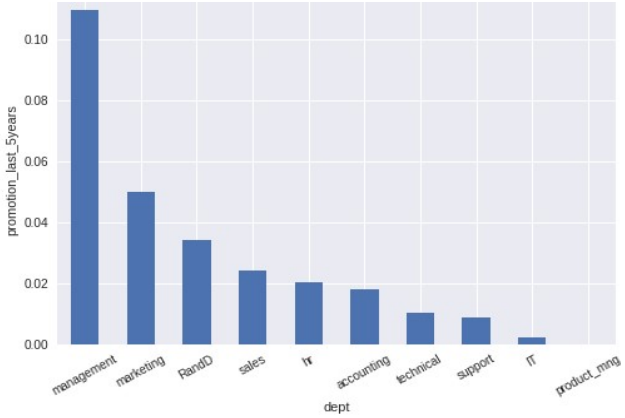


Figure 2: Promotion last 5 years

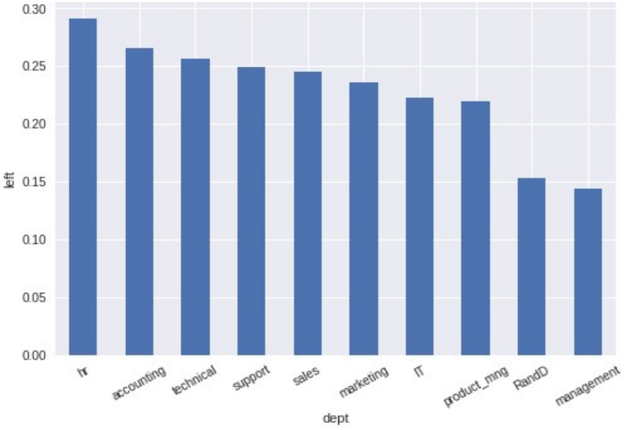


Figure 3 : Number of people who left their jobs

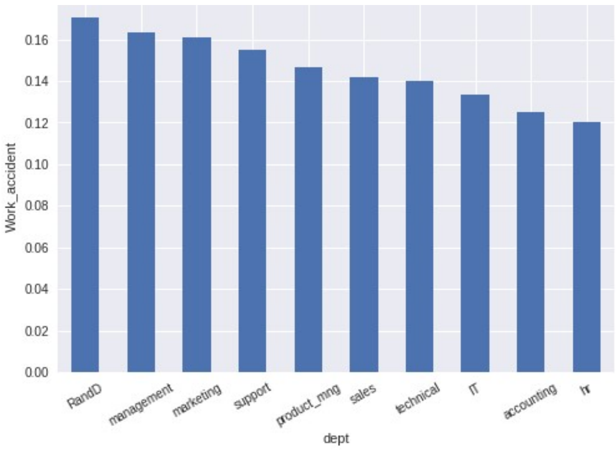


Figure 4 : Number of Accidents per department

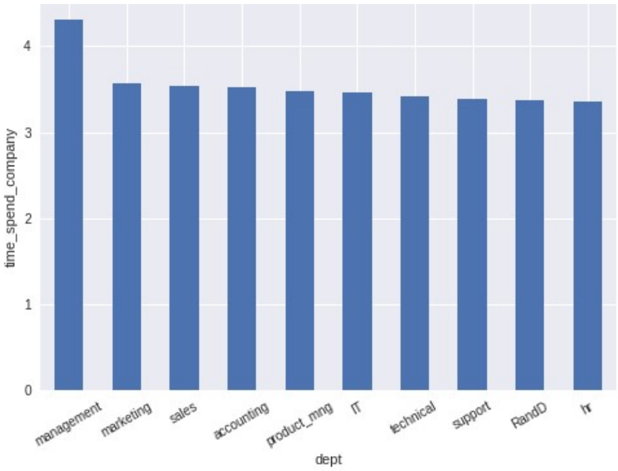


Figure 5 : Average time spent at a company

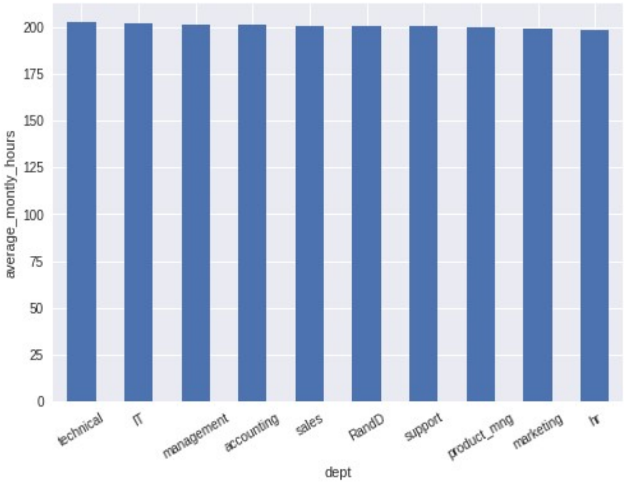


Figure 6 : Average Monthly hours

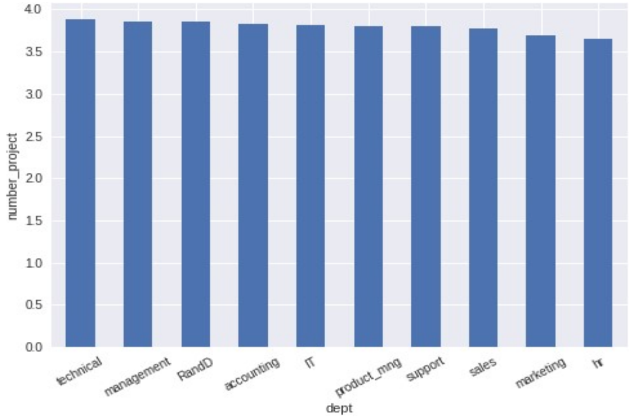


Figure 7 : Number of projects

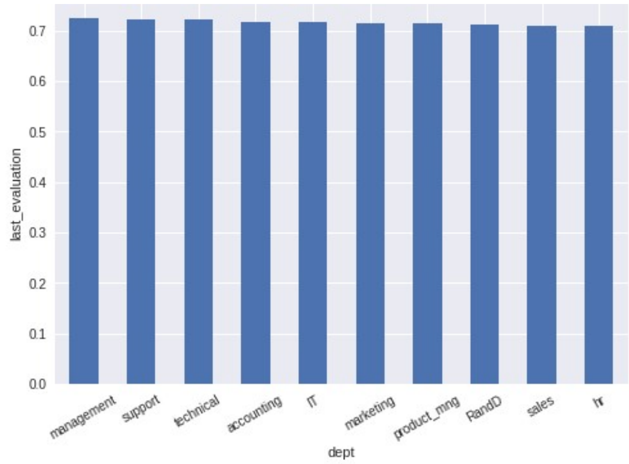


Figure 8: Last Evaluation Score

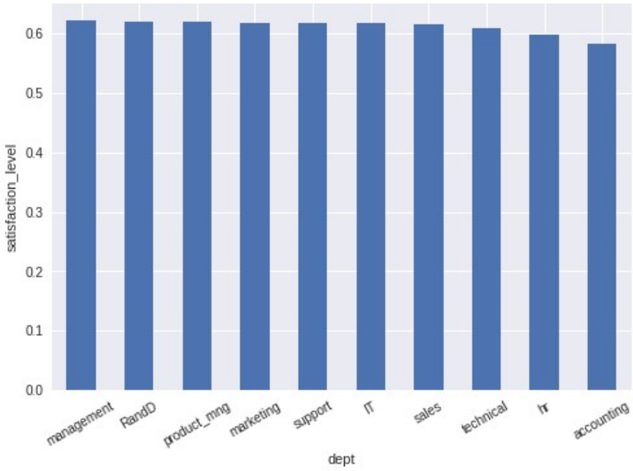
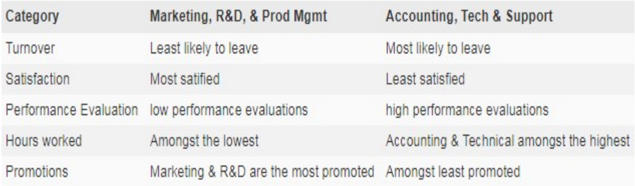


Figure 9 : Satisfaction Score

The above bar graphs give us more insights into the working trends by department. Although HR department itself has high turnover, we have omitted it from further analysis because their department is closely tied to recruitment and people management, so the high turnover in their department may actually be due to the high turnover in other departments.



Category	Marketing, R&D, & Prod Mgmt	Accounting, Tech & Support
Turnover	Least likely to leave	Most likely to leave
Satisfaction	Most satified	Least satisfied
Performance Evaluation	low performance evaluations	high performance evaluations
Hours worked	Amongst the lowest	Accounting & Technical amongst the highest
Promotions	Marketing & R&D are the most promoted	Amongst least promoted

Table 3 : Bar Graphs Analysis

Accounting, Tech and support have some of the hardest working people as far as hours, and they generally have higher performance evaluations, yet they are amongst the least satisfied, and least promoted. Marketing, R&D, & Prod Mgmt. generally work less as far as hours, do not do so well on evaluations, yet they are amongst the most satisfied, and most likely to get promoted. Promotion to management results in a large boost in pay, so it is easy to see why not being promoted, would result in leaving the company. While, the name of the company is undisclosed, I would guess that this company is largely dependent on Marketing and R&D, which is why those folks tend to get promoted more easily. This may be perceived as unjust by good employees of other departments, which is resulting in the high turnover by good employees.

Next, we will apply machine learning techniques, in order to understand which models work best in predicting which employees will leave next.

Logistic Regression

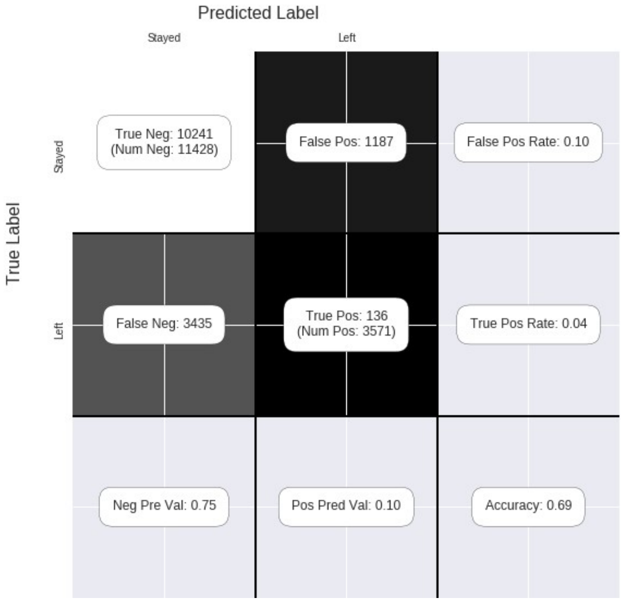


Figure 10: Regression model Results

Figure 10 above, this Logistic Regression model is only yielding a 70% accuracy. False Positive Rate is low at just 9%, which is good. True Positive Rate is also low at 3%, which is bad. This model is not very good at predicting who is likely to leave.

Random Forest

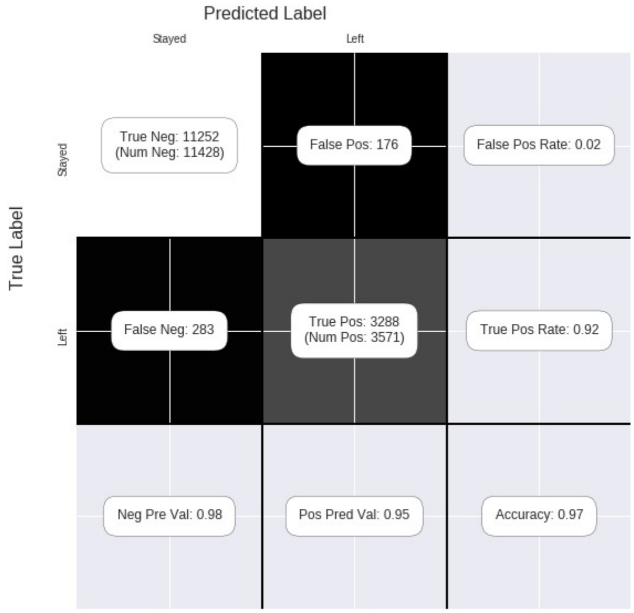


Figure 11 : Random Forest Classifier Results

A Random Forest Model is a better model with a 97% accuracy It is able to maintain a low False Positive Rate at 2% True Positive Rate is high at 93% Overall, this model does a better job is predicating which employees will leave next.

CONCLUSION

The model suggested in this study was the combination with Random Forest and Regression models. Features with higher classification effectiveness are more important and relevant for the specific classification task. The set of these important and relevant features is thus considered as the search starting point and is expected to have high relevance to the best feature subset for feature subset selection. The search starting point for feature subset selection is determined by somewhere in the middle of the search space. The results showed that the Random Classifier model used in this study could be the best model of categorizing, and the accuracy was 97%. The results showed that the best model of turnover prediction, this model could help individual industry to establish their database to investigate which factors could be used as prediction for employees' turnover. Because, there may be some signs before an employee really apply for turnover, the key point is that whether manager could notice or not. Consequently, this system is suggested for the industry to establish their own system to predict employees' turnover, and if the company wants to retain its valuable employers, it has to reduce/increase one of the factors in the efficiency. If so, either they have to

Increase no of projects or decrease average monthly hours, which is kind of impractical if both done at same time, Thus HR departments should adopt these Data Mining Techniques to enhance their decision making processes and be able to plan well in time for future problems that might arise due to stuff leaving.

References

[1] K. A. Nigam, K. S. Mccallum, Thrun, and T. Mitchell, “Text Classification from Labeled and Unlabeled Documents Using EM”, Machine Learning, vol. 39, 2000, pp. 103-134.

[2] J. G. Dy, and C.E. Brodley, “Feature Subset Selection and Order Identification for Unsupervised Learning”, in Proc. 17th Int’l Conf. Machine Learning, 2000, pp. 247-254.

[3] Carlo Dell' aquila, Francesco Di Tria, Ezio Lefons, Filippo Tangorra, "Business Intelligence Systems: A Comparative Analysis", Wseas Transactions on Information Science and Applications, Issue 5, Vol. 5, May 2008.

[4] F. W. Famili, M. Shen, R. Weber, and E. Simoudis, “Data Preprocessing and Intelligent Data Analysis”, Intell. Data Anal. 1(1-4), 1997, pp. 3-23.

[5] F.E. Emery, and E.C. Trist, “The Causal Texture of organizations”. Human Relations, vol. 18, 1965, pp. 21-31.

[6] J. L. Price, The study of turnover, Ames: Iowa State University Press, 1997.

[7] A. C. Bluedron, “The theories of turnover: Causes effects and eaning”, Research in the Sociology of Organization, vol.35, 1982, pp. 135-153.

[8] G. H. Ferguson, and W.F. Ferguson, “Distinguishing Voluntary from Involuntary, Nurse Turnover”, Nursing management, 17(12), 1986, pp. 43-44.

[9] J. E. Newman, “Predicting absenteeism and turnover: A field comparison of fishbein’s model and traditional job attitude measure”,, even if they have been submitted for publication, should be cited as “unpublished”

[10] M. A. Hall, “Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning”, in Proc. 17th Int’l Conf. Machine Learning, 2000, pp. 359-366.

[11] L. Yu, and H. Liu, “Feature Selection for HighDimensional Data: A Fast Correlation-Based Filter Solution”, In Proc. 20th Int’l Conf. Machine Learning, 2003, pp. 856-863.

[12] H. Liu , and Motoda, H. Feature Selection for Knowledge Discovery and Data Mining. Boston: Kluwer Academic,1998.

[13] R. O. Duda, and P. E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, 1973Writer’s Handbook. Mill Valley, CA: University Science, 1989.